KERJA PRAKTIK - KI141330

Topic Modelling pada Data Artikel Peneliti Penerima Dana PDUPT Menggunakan *Gensim*

Jurusan Teknik Informatika Fakultas Teknologi Elektro dan Informatika Cerdas Institut Teknologi Sepuluh Nopember
Jl. Teknik Kimia , Kec. Sukolilo, Kota SBY, Jawa Timur 60117

Periode: 1 Juni 2020 - 28 Agustus 2020

Oleh:
Bobbi Aditya 05111740000099

Pembimbing Jurusan
Dr. Diana Purwitasari, S.Kom, M.Sc.
Pembimbing Lapangan
Bu Esther Irawati Setiawan (iSTTS)

DEPARTEMEN INFORMATIKA
Fakultas Teknologi Elektro dan Informatika Cerdas
Institut Teknologi Sepuluh Nopember
Surabaya 2020

[Halaman ini sengaja dikosongkan]

KERJA PRAKTIK – KI141330

Topic Modelling pada Data Artikel Peneliti Penerima Dana PDUPT Menggunakan Gensim

Jurusan Teknik Informatika Fakultas Teknologi Elektro dan Informatika Cerdas Institut Teknologi Sepuluh Nopember
Jl. Teknik Kimia, Kec. Sukolilo, Kota SBY, Jawa Timur 60117

Periode: 1 Juni 2020 – 28 Agustus 2020

Oleh:
Bobbi Aditya 05111740000099

Pembimbing Jurusan
Dr. Diana Purwitasari, S.Kom, M.Sc.
Pembimbing Lapangan
Bu Esther Irawati Setiawan (iSTTS)

DEPARTEMEN INFORMATIKA
Fakultas Teknologi Elektro dan Informatika Cerdas
Institut Teknologi Sepuluh Nopember
Surabaya 2020

[Halaman ini sengaja dikosongkan]

Lembar Pengesahan
Kerja Praktik

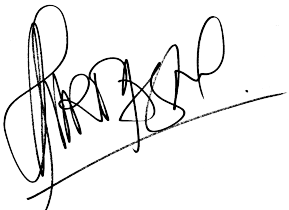
Topic Modelling pada Data Artikel Peneliti Penerima Dana PDUPT Menggunakan Gensim

Oleh:

Bobbi Aditya 05111740000099

Disetujui oleh Pembimbing Kerja Praktik:

1. Dr. Diana Purwitasari, S.Kom, M.Sc.
   NIP : 197804102003122001



(Pembimbing Departemen)

2. Dr. Esther Irawati Setiawan, S.Kom, M.Kom
   Dosen Teknik Informatika, Institut Sains dan Teknologi Terpadu Surabaya (ISTTS)
   Tim Riset PDUPT Kontrak No. 1218/PKS/ITS/2020

(Pembimbing Lapangan)

[Halaman ini sengaja dikosongkan]

Topic Modelling pada Data Artikel Peneliti Penerima Dana PDUPT Menggunakan Gensim

Nama Mahasiswa : Bobbi Aditya
NRP : 05111740000099
Departemen : Informatika FTEIC-ITS
Pembimbing Jurusan : Diana Purwitasari, S.Kom., M.Sc.

Abstrak

Penelitian merupakan tulang punggung dari kemajuan teknologi dan pendidikan. Indonesia sebagai negara berkembang dan negara dengan jumlah penduduk terbesar ke-4 di dunia, memiliki sumber daya manusia yang sangat besar. Dengan sumber daya yang sangat besar, penelitian di Indonesia seharusnya dapat memberikan pengaruh yang lebih banyak lagi. Indonesia masih belum memiliki penelitian kolaborasi antar peneliti yang cukup baik. Untuk itu, dalam penelitian ini diangkat sebuah model pengelompokan penelitian berdasarkan judul penelitian yang harapannya dapat menjadi batu loncatan awal dalam pembuatan sistem rekomendasi judul penelitian berdasarkan kecenderungan publikasi yang dilakukan oleh seorang peneliti. Metode pembuatan model yang dilakukan menggunakan LDA Mallet dengan melakukan pengelompokkan menjadi 18 kelompok topik. Hasil penelitian ini menghasilkan *conherence score* sebesar 0.56. Hasil tersebut menunjukkan bahwa model sudah cukup baik untuk digunakan dalam penelitian selanjutnya.

Kata Kunci : Penelitian, LDA Mallet,

KATA PENGANTAR

Puji dan syukur penulis panjatkan kepada Tuhan Yang Maha Esa karena atas berkat limpahan rahmat dan lindungan-Nya penulis dapat melaksanakan salah satu kewajiban sebagai mahasiswa Teknik Informatika ITS yaitu Kerja Praktik (KP).

Penulis menyadari masih terdapat banyak kekurangan baik dalam pelaksanaan kerja praktik maupun penyusunan buku laporan ini, namun penulis berharap buku laporan ini dapat menambah wawasan pembaca dan dapat menjadi sumber referensi. Penulis mengharapkan kritik dan saran yang membangun untuk kesempurnaan penulisan buku laporan ini.

Melalui laporan ini penulis juga ingin menyampaikan rasa terima kasih kepada kepada orang-orang yang telah membantu dalam pelaksanaan kerja praktik hingga penyusunan laporan Kerja praktik baik secara langsung maupun tidak langsung. Orang-orang tersebut antara lain adalah:

1. Orang tua penulis
2. Ibu Diana Purwitasari, S.Kom., M.Sc. selaku dosen pembimbing kerja praktik yang telah membimbing penulis selama kerja praktik berlangsung.
3. Ibu Dini Adni Navastara, S.Kom, M.Sc., selaku dosen yang telah membimbing penulis selama kerja praktik berlangsung.
4. Ibu Esther Irawati Setiawan selaku pembimbing lapangan selama kerja praktik yang telah memberikan bimbingan serta ilmunya kepada penulis.
5. Bapak Ary Mazharuddin, PhD selaku koordinator Kerja Praktik.

Malang, November 2020
Penulis

[Halaman ini sengaja dikosongkan]

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Penelitian merupakan tulang punggung dari kemajuan teknologi dan pendidikan. Indonesia sebagai negara berkembang dan negara dengan jumlah penduduk terbesar ke-4 di dunia, memiliki sumber daya manusia yang sangat besar. Dengan sumber daya yang sangat besar, penelitian di Indonesia seharusnya dapat memberikan pengaruh yang lebih banyak lagi.

Indonesia masih belum memiliki *platform* pencarian penelitian yang independen, para peneliti di Indonesia masih belum memiliki alat untuk mencari peneliti-peneliti lainnya yang memiliki kemiripan dengan penelitian yang dilakukannya. Kolaborasi antar peneliti di Indonesia masih susah untuk diwujudkan.

Untuk itu, dalam penelitian ini diangkat sebuah model pengelompokan penelitian berdasarkan judul penelitian yang harapannya dapat menjadi batu loncatan awal dalam pembuatan sistem rekomendasi judul penelitian berdasarkan kecenderungan publikasi yang dilakukan oleh seorang peneliti. Dengan terciptanya sistem rekomendasi tersebut, kolaborasi antar peneliti di Indonesia diharapkan bias semakin meningkat serta penelitian dan pendidikan di Indonesia dapat semakin maju.

1.2 Tujuan

Tujuan diadakan penelitian ini adalah untuk membuat pengelompokkan judul penelitian berdasarkan topik penelitian, serta melakukan analisa hasil kelompok topik penelitian yang terbentuk

1.3 Manfaat

Hasil penelitian ini diharapkan dapat bermanfaat bagi beberapa pihak.

Manfaat untuk peneliti :

- Peneliti dapat menggunakan model yang ada untuk melihat penelitian yang dilakukan tergolongkan dalam topik apa

Manfaat untuk penelitian selanjutnya:

- Model yang terbentuk dapat digunakan sebagai dasar dalam memberikan rekomendasi judul penelitian berdasarkan kemiripan judul

1.4 Rumusan Masalah

Rumusan masalah yang diangkat dalam penelitian ini adalah, bagaimana cara membuat pengelompokkan judul penelitian berdasarkan topik penelitian yang baik?

1.5 Lokasi dan Waktu Kerja Praktik

Kerja praktik kali ini dilaksanakan pada waktu dan tempat sebagai berikut:

Lokasi : Departemen Teknik Informatika ITS

Alamat : Jl. Teknik Kimia, Institut Teknologi Sepuluh November, Jawa Timur 60117

Waktu : 1 Juni 2020 – 28 Agustus 2020

1.6 Metodologi Kerja Praktik

1. **Perumusan Masalah**
   Untuk memahami permasalahan yang ingin diangkat, pembimbing menjelaskan mengenai *problem* yang ada serta memberikan gambaran tentang penyebab diangkatnya masalah ini. Kemudian dilakukan diskusi lebih lanjut untuk membahas hal yang diperlukan agar dapat dipelajari terlebih dahulu sebelum melakukan penelitian.
2. **Studi Literatur**
   Pada tahap studi literatur, penulis melakukan eksplorasi dan mempelajari berbagai materi yang dapat mendukung proses penelitian. Eskplorasi diutamakan pada *topic modelling* serta library *gensim*.
3. **Pembuatan Model**
   Pembuatan model yang dilakukan meliputi pra-proses data awal hingga terbentuknya sebuah model yang dapat mengelompokkan judul penelitian dengan baik.
4. **Pengujian dan Evaluasi**
   Pengujian dilakukan terhadap model yang terbntuk dengan melakukan beberapa skenario pengujian serta melakukan analisa kelayakan terhadap hasil pengelompokkan yang terbentuk.

1.7 Sistematika Laporan

Laporan kerja praktik ini terdiri dari 5 bab dengan rincian sebagai berikut:

1. **Bab I: Pendahuluan**
   Pada bab ini dijelaskan latar belakang masalah, tujuan, manfaat, rumusan masalah, lokasi dan waktu kerja

praktik, metodologi kerja praktik, dan sistematika laporan.

2. **Bab II: Tinjauan Pustaka**
   Pada bab ini dijelaskan dasar teori dan teknolgo yang dipakai dalam pembuatan model
3. **Bab III: Metodologi Penelitian**
   Pada bab ini dijelaskan metodologi penelitian yang akan dilakukan
4. **Bab IV: Implementasi dan Analisa**
   Pada bab ini akan dibahas tentang proses pembuatan model serta analisa yang dilakukan terhadap model yang terbentuk
5. **Bab V: Kesimpulan dan Saran**
   Pada bab ini dijelaskan mengenai kesimpulan dan saran yang didapatkan dari penelitian yang sudah dilakukan

BAB 2
TINJAUAN PUSTAKA

2.1 PDUPT

PDUPT adalah penelitian yang mengacu pada bidang unggulan yang telah ditetapkan dalam Rencana Strategis Penelitian Perguruan Tinggi. Penelitian ini terarah dan bersifat *top-down* atau *bottom-up* dengan dukungan dana, sarana dan prasarana penelitian dari perguruan tinggi serta *stakeholders*. [1]

PDUPT ditujukan untuk mencapai penelitian unggulan perguruan tinggi yang pada tahapan model/produk/purwarupa yang telah di ujicoba dalam lingkungan yang sebenarnya.

2.2 *Google Cendekia*

Google Cendekia menyediakan layanan sederhana untuk mencari literatur ilmiah secara luas. Dari satu tempat, kita dapat menelusuri berbagai disiplin ilmu dan sumber: artikel, tesis, buku, abstrak, dan opini pengadilan, dari penerbit akademis, perkumpulan profesional, repositori online, universitas, dan situs web lainnya. *Google Cendekia* membantu kita menemukan penelitian yang relevan dari penelitian ilmiah yang dilakukan di seluruh dunia. [2]

2.3 Pra-Proses Data

2.3.1 *Tokenization*

Tokenization adalah sebuah proses memisahkan sebuah frasa, kalimat, paragraph atau sebuah teks dokumen yang utuh menjadi kumpulan kata/*term*. Setiap kata yang terbentuk disebut sebagai *token*. [3]

Proses ini adalah salah satu proses penting dalam serangkaian proses NLP karena makna sebuah kalimat/sebuah text dapat diawakilkan dengan menganalisa masing-masing kata yang mewakili kalimat/text tersebut.

2.3.2 *Stop Words*

Stop words adalah kata umum yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki makna. Stop words umumnya dimanfaatkan dalam *task information retrieval*. [4]

2.3.3 *N-Grams*

N-gram merupakan salah satu proses yang secara luas digunakan dalam *text mining* (pengolahan teks) dan pengolahan bahasa. N-gram adalah potongan N-karakter yang diambil dari suatu string. [5]

2.3.4 *Dictionary*

Dictionary adalah proses pengubahan kata menjadi sebuah id unik. *Dictionary* mengubah kata dalam dokumen menjadi representasi angka yang dapat dikenali dan diproses dengan lebih mudah. Proses pembuatan *dictionary* dapat dibantu dengan menggunakan *library genism*. [6]

2.3.5 *Corpus*

Corpus adalah sekumpulan *object document*. Setiap dokumen akan diubah menjadi representasi angka berdasarkan *dictionary* yang sudah dibentuk. *Corpus* digunakan sebagai *input* dari proses *training* sebuah *model*. Selama proses training, *model* menggunakan *corpus* untuk melihat kesamaan topik dan inisialisasi *model* parameter. [6]

2.4 *Dunn Index*

Dunn index adalah sebuah metrik untuk melakukan evaluasi terhadap hasil *clustering/*pengelompokkan. Perhitungan Dunn Index yang dilakukan pada penelitian ini dihitung berdasarkan rata-rata *cosine similarity* dari sebuah judul terhadap judul lainnya pada sebuah kelompok topik.

2.5 *t-Distributed Stochastic Neighbor Embedding*

t-Distributed Stochastic Neighbor Embedding (t-SNE) adalah sebuah algoritma untuk melakukan reduksi dimensi yang dikembangkan oleh Laurens van der Maaten and Geoffrey Hinton pada tahun 2008.

t-SNE melakukan perhitungan similaritas antara pasangan atribut pada dimensi tinggi dan pada dimensi rendah. Algoritma ini kemudian melakukan optimasi terhadap perhitungan similaritas ini menggunakan cost function. [7]

Tujuan dilakukan t-SNE pada umumnya untuk melakukan visualisasi data. Data dengan jumlah atribut yang banyak (dimensi tinggi) bisa di wakilkan oleh 2 atribut (dimensi rendah). Dengan dimensi yang rendah ini, data akan bisa digambarkan di ruang 2 dimensi

BAB 3
METODOLOGI PENELITIAN

3.1 Sumber Data

Data yang digunakan adalah judul penelitian/artikel pada website *Google Cendekia* dari para peneliti yang tergolong dalam penelitian PDUPT 3 tahun terakhir sejumlah **181.326** judul dan **3.894** peneliti.

3.2 Struktur Data

Struktur data yang akan digunakan pada penelitian dapat dilihat pada tabel 1

Tabel 1 - Struktur Data

No.	F
1	y1
2	y2
3	y3
.	.
.	.
.	.
n	yn

Untuk setiap data Judul (y) akan memiliki data judul (F) berupa kalimat. Kalimat judul ini yang akan digunakan dalam proses analisa dan pembuatan model

3.3 Langkah Analisis

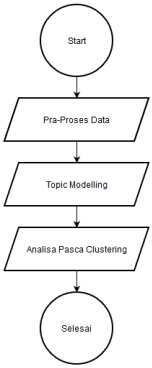
Langkah analisis yang dilakukan pada penelitian ini adalah sebagai berikut:

1. Melakukan pra proses data yang teridiri dari proses:
    a. Membersihkan judul dari karakter-karakter kata yang dapat mengganggu proses *modelling*.
    b. Melakukan *tokenizing*.

c. Melakukan proses *lemmatization* untuk masing-masing token kata.
d. Melakukan proses *stemming* pada masing-masing token kata.
e. Membuat *dictionary* dan *corpus* yang akan digunakan dalam proses pembuatan model.

2. Pembuatan model untuk mengelompokan judul menjadi n topik.
3. Melakukan analisa terhadap kelompok topik yang terbentuk dengan cara:
    a. Melakukan *manual labelling* terhadap kelompok topik yang terbentuk.
    b. Menghitung *dunn index* hasil pengelompkan judul.
    c. Melihat perseberan hasil pengelompokan judul menggunakan t-SNE.
    d. Melakukan analisa kedekatan judul pada kelompok judul yang berdekatan.

Pada gambar 1 dapat dilihat diagram alir dari langkah analisis yang dilakukan



Gambar 1 – Flow Chart

BAB 4
IMPLEMENTASI DAN ANALISA

4.1 Pra-Proses Data

Pada tahap pra-proses, data akan dibersihkan dan diubah bentuknya agar siap untuk dimasukkan ke dalam model. Data judul akan diubah menjadi data numerik yang dapat mewakili masing masing judul yang ada.

4.1.1 Pembersihan data

Pada data judul yang ada, masih ditemukan beberapa karakter non *alphabetic* dan beberapa karakter lainnya yang mengganggu. Untuk mengatasi hal tersebut dilakukan proses pembersihan data dimana akan diambil karakter yang tergolong *alphabetic*. Selain itu juga dilakukan *lower-case* kalimat agar keseluruhan data menjadi seragam.

4.1.2 *Tokenizing*

Setelah melakukan pembersihan data, data yang sudah dibersihkan selanjutnya akan di ubah bentuknya menjadi token.

Pada proses tokenizing dibagi menjadi 2 bagian, yakni proses *stop words removal* serta pembuatan *bigram.*

4.1.2.1 Stop Words Removal

Proses ini dilakukan untuk melakukan penghapusan kata/token yang tergolong dalam *stop words*. List *Stop Words* diambil dari dua Bahasa (Bahasa Inggris dan Bahasa Indonesia) menyesuaikan Bahasa pada data judul yang ada. Untuk list stop words ini diambil dari library NLTK, serta ditambahkan beberapa

kata dari hasil analisa.Total jumlah *stop words* yang digunakan adalah 967 kata.

Tabel 2 – Tambahan Stop Words

Tambahan Stop Words			
of	to	pro	hasil
in	by	analysis	
and	as	berbasis	
the	and	tahun	
for	an	between	
on	pengaruh	kualitas	
using	effect	method	
based	analisis	metode	
from	at	through	
with	pre	menggunakan	

4.1.2.2 Pembuatan *Bigram*

Proses ini dilakukan untuk menggubah kata yang tergolong unigram menjadi bigram menggunakan *library* gensim. Hal ini dilakukan untuk membuat kata-kata yang sering muncul bersamaan agar menjadi 1 token yang sama, sehingga dikenali menjadi 1 makna yang sama.

4.1.3 *Lemmatization*

Proses *lemmatization* dilakukan untuk mengubah sebuah kata menjadi bentuk dasar dari kata tersebut dengan melihat jenis kata. Proses ini difokuskan pada data judul yang berbahasa inggris. Proses ini dilakukan dengan bantuan dari library spacy dengan model yang digunakan adalah model *'en'* .

4.1.4 *Stemming*

Proses *stemming* dilakukan untuk mengubah sebuah kata menjadi bentuk dasar dari kata tersebut tanpa melihat jenis kata. Proses ini difokuskan pada data judul yang berbahasa Indonesia. Proses ini dilakukan dengan bantuan *StemmerFactory* dari library Sastrawi

4.1.5 Pembuatan Dictionary & Corpus

Proses pembuatan *dictionary* dilakukan untuk mengubah kata menjadi sebuah angka. Setelah masing-masing kata berubah menjadi angka, setiap judul akan diubah menjadi representasi numerik berdasarkan *dictionary* yang ada dalam sebuah *corpus*. Masing-masing kalimat akan diubah menjadi list id kata beserta frekuensi kemunculan kata tersebut. Jumlah kata dalam dictionary yang terbentuk **1.632.911.**

4.2 **Topic Modelling**

Dalam pembuatan model pengelompokkan, dilakukan beberapa proses untuk mendapatkan model terbaik.

4.2.1 Pemilihan model

Topic Modelling dilakukan dengan menggunakan model *LDA Mallet*.

4.2.2 *Hyperparameter Tuning*

Tahap hyperparameter tuning dilakukan untuk memiliih parameter yang dapat mengahasilkan *conherence score* terbaik. Pada tabel 3 dapat dilihat *hyperparameter* yang akan diujikan.

Tabel 3 – Hyperparameter Tuning

n Topik	Conherence Score
10	0.49
30	0.51
50	0.5
70	0.48
90	0.472
18	0.56
26	0.54
34	0.51
42	0.52

Pada proses hyperparameter tuning didapatkan model ***LDA Mallet* dengan 18 topik** menghasilkan *conherence score* terbaik dengan nilai **0.56**

4.2.3 Evaluasi Model

Untuk melakukan evaluasi model lanjutan, dilakukan perhitungan *dunn index* pada model yang terbentuk. Pada tabel 4 dapat dilihat uji coba serta nilai *dunn index* yang dihasilkan.

Tabel 4 - Uji Coba Dunn Index

Model	Jumlah Topik	Dunn Index
Lda Mallet	18	0.001
	30	0.0001
	50	0.00015
LSI	18	0.001
	30	-
	50	-

Pada tabel 5 dapat dilihat detil dari perhitungan dunn index percobaan dengan jumlah *n_topic* 18 topik.

Tabel 5 – Dunn Index per Topik

Topic	Avg Cosine Sim	
	Lda Mallet 18 Topic	LSI 18 Topic
0	0.017	Error
1	0.029	Error
2	0.011	0.095
3	0.017	0.127
4	0.012	0.048
5	0.011	0.019
6	0.024	Error
7	0.007	0.021
8	0.015	0.082
9	0.017	0.077
10	0.018	0.06
11	0.011	0.019
12	0.027	0.018
13	0.022	0.095
14	0.012	0.015
15	0.013	0.036
16	0.011	0.01
17	0.015	0.015
Dunn Index	**0.001**	0.001

Model LDA Mallet dengan 18 topik memiliki nilai dunn index terbaik dengan nilai **0.001**

4.3 **Analisa Pasca Clustering**

4.3.1 *Manual Labelling*

Berdasarkan data masing-masing judul yang sudah dikelompokkan berdasarkan model, dilihat konteks masing-masing judul dalam 1 kelompok topik secara manual untuk pembuatan label. Pada tabel 6 dapat dilihat hasil *manual labelling* yang berhasil dilakukan.

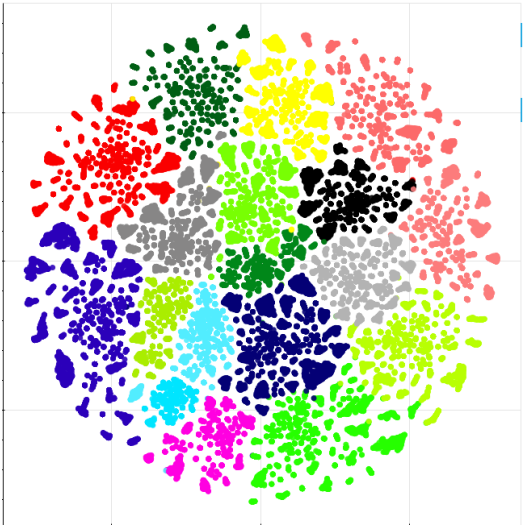
Tabel 6 – Topik Manual Labelling

No	Topik	No	Topik
1	Kimia	10	Pendidikan Sains
2	Hukum	11	HRD
3	Nutrisi-Bioteknologi	12	Peternakan
4	Pertanian	13	Machine Learning
5	Mikrobiologi	14	Konservasi
6	Electrical Engineering	15	Ekonomi Pemberdayaan Masyarakat
7	Pendidikan	16	Kimia Organik
8	Lingkungan	17	Penyakit
9	Elektro-Mesin	18	Kesehatan Masyarakat

4.3.2 t-SNE

t-SNE dilakukan untuk membuat sebuah nilai yang dapat merpresentasikan judul agar dapat dalam sebuah grafik/*plot*. Fitur yang akan digunakan dalam t-SNE adalah nilai probabilitas 18 topik masing-masing judul, sehingga masing-masing judul akan memiliki 18 fitur yang akan direduksi menjadi 2 fitur menggunakan t-SNE.

Data yang digunakan dalam proses ini adalah data yang sudah direduksi berdasarkan nilai probabilitas judul pada topik utama judul tersebut. Hal ini agar dapat melihat representasi terbaik dari masing-masing topik.



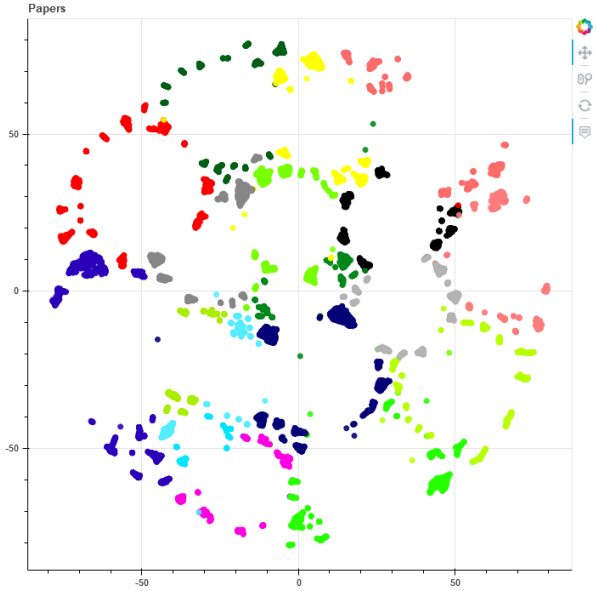
Gambar 2 – Hasil t-Sne

Dari hasil *plot* t-SNE pada gambar 2 dapat dilihat bahwa masing-masing topik sudah terkelompokkan dengan baik.

4.3.3 Analisa Kedekatan Judul

Analisa kedekatan judul dilihat berdasarkan hasil *plot* t-SNE, dari hasil yang ada dapat dilihat bahwa terdapat judul-judul yang berdekatan pada daerah pinggiran *cluster*. Asumsinya semakin dekat sebuah titik dengan titik lainnya, maka akan semakin mirip konteks judul dari artikel tersebut. Untuk itu dilakukan observasi daerah pinggiran cluster.

Proses pengambilan daerah pinggiran *cluster* dilakukan dengan cara menghitung jarak masing-masing titik dengan *centroid* sebuah *cluster*. Dari data jarak ini kemudian diambil data artikel yang termasuk ke dalam persentil 75% (dengan kata lain artikel yang jauh dari *centroid*). Pada gambar 3 dapat dilihat daerah pinggirian cluster yang dianalisa.



Gambar 3 – Daerah Pinggiran Cluster

Dari data pinggirian ini kemudian dilakukan analisa untuk menguji asumsi yang dimiliki. Untuk itu dilakukan perhitungan cosine similarity dan distance antar judul yang berada pada perbatasan topik. Dari 18 topik, didapatkan **39 kombinasi daerah pinggiran.** Analisa dilakukan berdasarkan distance, cosine sim dan konteks dari masing-masing judul artikel yang ada.

Berikut adalah beberapa poin yang berhasil diambil dari analisa daerah pinggiran:

- Tidak semua daerah pinggiran memiliki konteks yang sama
- Jarak antar daerah pinggiran sebenarnya cukup jauh apabila dilihat lebih dekat lagi
- Titik-titik yang sangat dekat apabila memliki topik yang berbeda sebenarnya memiliki konteks yang hampir sama
- Kesamaan konteks juga dipengaruhi oleh korelasi dari topik besar judul

BAB 5
KESMIMPULAN & SARAN

5.1 Kesimpulan

Dari percobaan dan pembuatan model yang dilakukan model yang didapatkan sudah cukup baik dalam melakukan pengelompokkan judul berdasarkan topik. Dengan melakukan pengelompokkan menjadi **18 topik** dengan **nilai *conherence score 0.56*** model dapat digunakan dalam penelitian selanjutnya

5.2 Saran

Dari percobaan yang dilakukan masih terdapat banyak kekurangan dan bagian yang masih bisa dikembangkan lagi, berikut adalah bagian-bagian yang masih bisa dikembangkan:

- Judul masih dalam bentuk bahasa campuran (Inggris, Indonesia ataupun bukan keduanya). Hal ini menyebabkan deteksi topik jadi bias terhadap Bahasa juga, akan lebih baik hasilnya apabila melakukan *modelling* terhadap judul judul artikel dalam 1 bahasa yang sama
- *Stop words* masih bisa ditambah lebih banyak lagi
- Masih terdapat beberapa judul yang sama pada data
- Pra-proses masih bisa diperbaiki lagi (menghilangkan karakter-karakter yang tidak dipakai, proses stemming dan lemma menyesuaikan Bahasa artikel)
- Proses n-gram bisa di uji coba dengan n yang lebih banyak

DAFTAR PUSTAKA

[1] Ristek Dikti (2016). Slide *powerpoint* Penelitian Dasar Unggulan Perguruan Tinggi (PDUPT)

[2] Google (2020). Tentang Google Cendekia [online]. Available at https://scholar.google.co.id/intl/id/scholar/about.html [Accessed 31 October 2020]

[3] Shuban Singh(2019). How to Get Started with NLP – 6 Unique Methods to Perform Tokenization [online]. Available at https://www.analyticsvidhya.com/blog/2019/07/how-get-started-nlp-6-unique-ways-perform-tokenization/ [Accessed 31 October 2020]

[4] Yudi Wibisono(2008). Stop Words untuk Bahasa Indonesia [online] . Available at https://yudiwbs.wordpress.com/2008/07/23/stop-words-untuk-bahasa-indonesia/ [Accessed 31 October 2020]

[5] Kavita Ganesan(2014). What are N-Grams [online]. Available at https://kavita-ganesan.com/what-are-n-grams/#.X5-KjFAxWMp [Accesed 31 October 2020]

[6] Radim Rehurek(2020) Gensim [Online]. Available at https://radimrehurek.com/gensim/ [Accessed 31 October 2020]

[7] Violante, A. (2018). An Introduction to t-SNE with Python Example [online]. Available at https://towardsdatascience.com/an-introduction-to-t-sne-with-python-example-5a3a293108d1 [Accessed 31 October 2020]

BIODATA PENULIS

Bobbi Aditya, lahir pada tanggal 18 Desember 1999 di Malang. Penulis merupakan mahasiswa yang sedang menempuh studi di Departemen Informatika Institut Teknologi Sepuluh Nopember (ITS). Penulis aktif dalam berorganisasi di Himpunan Mahasiswa Teknik Computer-Informatika tahun 2019/2020 dalam departemen hubungan luar sebagai staff, serta penulis menjadi Badan Pengurus Harian 2 bagian Hubungan Masyarakat pada rangkaian acara Schematics 2019